Capacity Expansion with Lead Times and Correlated Random Demand

Sarah M. Ryan
Department of Industrial & Manufacturing Systems Engineering
Iowa State University
Ames, Iowa 50011-2164

Voice: 515-294-4347
Fax: 515-294-3524
smryan@iastate.edu

October, 2001

This is the peer reviewed version of the following article: Ryan, S. M., "Capacity Expansion with Lead Times and Correlated Random Demand," Naval Research Logistics, 50(2), pp. 167-183 (2003), which has been published in final form at http://dx.doi.org/10.1002/nav.10055. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Capacity Expansion with Lead Times and Correlated Random Demand

Abstract

The combination of uncertain demand and lead times for installing capacity creates the risk of shortage during the lead time, which may have serious consequences for a service provider with dependent customers. This paper analyzes a model of capacity expansion with correlated random demand and a fixed lead time for adding capacity. We develop and prove the form of an optimal policy for expansion timing and size. We study the effects of demand correlation and randomness as well as the lead time length on the policy parameters. Correlation acts similarly to randomness in hastening expansions but has a smaller impact than randomness, especially when lead times are short. However, the failure either to recognize correlation or to accurately estimate its extent can cause substantial policy errors.

1. Introduction

When demand for capacity is uncertain and significant lead times exist for adding capacity, managers must carefully consider when to initiate new capacity additions. Discounting future costs encourages the delay of capacity expansion to the latest possible moment. However, postponing capacity additions increases the risk of capacity shortage during the installation lead time.

This paper analyzes a capacity expansion model in which installation lead times are fixed and the only source of uncertainty is the demand for capacity. The demand process studied exhibits both autocorrelation (which causes the mean rate of growth to change over time) and randomness (as demand increments vary about their mean). Both of these characteristics contribute to the unpredictability of future demand, but in different ways. Our goals are to develop an expansion policy that takes these demand characteristics into account, study their effects on the policy parameters and cost, and examine the results of failing to account for them properly.

The model and parameter values studied in this paper were motivated originally by a situation encountered in the utilities division of a large chemical manufacturer. This division's primary responsibility is providing steam for process heat, in addition to electric power, refrigeration and water services. In contrast to electric power, steam cannot be purchased from an outside utility. Insufficient steam pressure can lead to product quality problems or, in extreme cases, a long, expensive and embarrassing forced shutdown of portions of the plant. Demand for steam capacity, when adjusted for seasonal variations, has been steadily increasing, in a trend that is expected to continue as the plant grows. Though demand is seasonal, capacity must be built to

handle peak demands. The forecasting model that best fits the historical peak data exhibits autocorrelation. Additional steam generating capacity can be provided by facilities ranging from small gas-fired boilers purchased essentially off the shelf and installed in a matter of months, to a large coal-fired boiler, which would require numerous and extensive environmental justifications and a long construction period. All recent and projected expansions consist of one or more gas fired boilers.

This study, assumes that capacity shortages are expensive and to be avoided as much as possible. The same situation is likely to occur in the communications industry, particularly for providers of services such as Internet connections and wireless communication. To maintain customer goodwill, it is essential to maintain enough capacity to meet the demand for service when it occurs. Customer frustration with being unable to access the service arises from the fact that in most cases they cannot switch immediately to a competing provider. However, repeated failures to connect will cause customers to transfer their subscriptions to a more reliable service provider.

An expansion policy specifies both the timing and sizes of expansions. We assume that the managerial goal is to minimize the infinite horizon expected discounted cost of meeting some specified service level, which is measured in terms of a maximum allowable capacity shortage during an expansion lead time. We develop and prove the optimal form for the timing policy under some assumptions about the measurement of shortages. Then, given this timing policy, we show that, consistent with previous models without lead times or demand correlation, it is optimal to always install the same quantity of capacity. The expansion size balances the opposing effects of discounting and economies of scale. Since expansion timing has a larger

impact on both potential shortages and expected expansion cost than expansion sizes, most of the analysis focuses on timing. Using simulation to estimate a more complicated measure of lead time shortage, we study the effects of demand correlation and randomness on the timing policy. Finally, we examine the effects on the expansion policy and its cost of either ignoring or incorrectly estimating the amount of correlation in the demand.

Several authors have studied capacity expansion problems under various types of uncertainty, assuming that new capacity additions are obtained instantaneously. If lead times are negligible, then the timing of capacity additions can follow the realization of demand growth. The earliest work therefore focused on the sizes of capacity additions. Manne [11] modeled uncertain demand as a Brownian motion with deterministic drift and showed how uncertainty prompts larger capacity additions. He also showed that if the backlog penalty is low, and it is optimal to allow regular backlogs in capacity, then both the optimal capacity increment and the still absolute value of the optimal backlog trigger level increase with uncertainty. That is, increased uncertainty in demand can lead to delays in optimal capacity additions. Giglio [8] studied expansion policies allowing backlog assuming a linear trend and various types and amounts of demand uncertainty. He focused mainly on timing and argued that when shortage costs are high, increasing uncertainty causes earlier expansions. Freidenfelds [7] modeled demand as a birth and death process and showed how to derive an equivalent (higher) deterministic demand. The uncertainty therefore increased the size of the optimal capacity addition and prompted earlier installations. Bean et al. [2] generalized the demand models of the previous two papers to either a transformation of Brownian motion with drift or a semi-Markovian birth and death process. Assuming a fixed set of possible discrete facility sizes, and allowing no backlog, they

showed that the effect of uncertainty is to lower the interest rate, so that capacity is added sooner than it would be under deterministic demand. Rocklin et al. [15] proved the optimality of a nonstationary (*s*,*S*) policy to minimize the sum of linear capital, labor, maintenance and underprovision costs in a finite horizon model with independent random demands.

Lead times have been treated in a few papers, either as decision variables or as fixed quantities. Nickell [14] assumed that the timing of future changes in demand, such as a jump in its constant level or a change in its rate of increase, was the only source of uncertainty. He showed that the existence of a fixed capacity delivery lead time would cause a firm to introduce capacity increases earlier, with a longer lead time resulting in earlier anticipation of demand increases. Davis et al. [6] modeled demand as a random point process and allowed for stochastic nonzero lead times that depended on the controllable rate of investment in new capacity. They then analyzed the capacity expansion model as a stochastic control problem and computed the optimal policy in some simple cases. Chaouch and Buzacott [5] assumed fixed lead times for installing manufacturing capacity and modeled demand as an alternating renewal process, consisting of alternating periods of linear growth in demand and constant demand. They showed how to find the optimal plant size as well as the optimal capacity surplus or deficit to trigger a new capacity addition. In numerical tests, with relatively small penalties for capacity deficits, they showed that longer lead times cause increases in both the optimal trigger levels and the optimal sizes of capacity additions. Angelus et al. [1] formulated a finite horizon capacity expansion model applicable to the semiconductor industry. Assuming fixed lead times and autocorrelated random demand, they proved the optimality of an (*s*,*S*) type policy, in which the expansion point (*s*) and the expansion level (*S*) depend on the current period and its observed demand. The effect of

correlation was not studied specifically. Çakanyıldırım and Roundy [4] provided an algorithm to compute optimal expansion times for semiconductor production capacity with fixed lead times for stochastically increasing demand over a finite horizon.

Most of the numerical examples in the previous papers assumed a relatively small shortage penalty. Ryan [16] performed an empirical study comparing the expansion timing decisions that result from hedging the demand forecast with its upper prediction limit (UPL) with those that are dictated by maintaining a fixed excess capacity buffer. At the expense of a small increase in discounted expansion cost, use of the UPL hedge significantly reduced the sizes and frequency of shortages. McAllister and Ryan [12] further showed that these results held not only on average, but for nearly every demand realization.

Section 2 outlines the model including the characteristics of the demand process under study. In Section 3 we develop and prove the form of an optimal policy, which includes a forecast-adjusted minimum threshold level of excess capacity that prompts an expansion. We study the effects of demand randomness and correlation on this threshold in Section 4. In Section 5 we analyze the qualitative effects on the policy parameters and quantitative effects on the cost of errors in specifying the demand process and estimating its parameters. Section 6 concludes the paper.

2. Capacity Expansion Model

In many situations demand may be expected to increase linearly with time but the slope of the linear process can fluctuate. Random variation also contributes to unpredictability. Let d_t be the

demand for product or service in period *t*. The time (epoch) *t* marks the beginning of period *t*+1. We assume that d_t follows an integrated moving average (IMA) process with drift, given by:

$$\begin{aligned} d_t &= d_{t-1} + \mu - (1-\lambda)\varepsilon_{t-1} + \varepsilon_t, \\ d_0 &= \varepsilon_0 = 0. \end{aligned} \tag{1.1}$$

Here, $\mu > 0$ represents the deterministic trend component, $0 \le 1-\lambda \le 1$ is the moving average parameter, and the random shocks $\{\varepsilon_t, t = 1, 2, \ldots\}$ are independent, identically distributed normal random variables with mean 0 and variance σ^2. This model provides the best fit to the historical deseasonalized steam demand experienced by the chemical manufacturer discussed in Section 1.

Note that since the random shocks can be negative, it is possible for demand to decrease. However, since $\mu > 0$, the general trend is increasing. Also, capacity expansion decisions are generally made according to the maximum demand observed so far, which is always increasing.

The impact of the value of λ is easiest to see by expressing the demand process in "random shock" form [3]. Recursively substituting for $d_{t-1}, d_{t-2}, \ldots,$ we obtain:

$$d_t = \mu t + \lambda \sum_{i=1}^{t-1} \varepsilon_i + \varepsilon_t \tag{1.2}$$

If $\lambda = 0$, this process is a simple uncorrelated linear trend process. At the other extreme, as $\lambda \to 1$, the demand process becomes increasingly autocorrelated and approaches a random walk with deterministic drift and variance σ^2. Graves [9] has recently studied an inventory model having a similar demand process without drift, where he termed it a nonstationary demand model. Because the term *nonstationary* has not had a consistent definition in the literature, in this paper we focus on autocorrelation as the distinguishing feature of the demand model.

The covariance of the demands in separate periods can be derived from Equation (1.2) as

$$\mathrm{Cov}\left[d_k, d_{k+j}\right] = \sigma^2\left(\lambda + \lambda^2(k-1)\right). \tag{1.3}$$

If $\lambda = 0$ then demands in different periods are uncorrelated. If $\lambda > 0$ then the correlation between the demands is given by

$$\mathrm{Corr}\left[d_k, d_{k+j}\right] = \frac{\lambda + \lambda^2(k-1)}{\sqrt{\lambda + \lambda^2(k-1)}\sqrt{\lambda + \lambda^2(k+j-1)}}, \tag{1.4}$$

which is close to one when j is small relative to k and $\lambda \neq 0$, and increases with λ. Therefore, we refer to λ as the correlation parameter. With a large value of λ, the IMA process with drift can model demand that has eras of rapid growth interspersed among eras of stagnant growth.

The random shock form is also most convenient for forecasting. Looking ahead k periods from the current period, τ, the demand is given by

$$d_{\tau+k} = d_\tau - (1-\lambda)\varepsilon_\tau + \mu k + \lambda\sum_{i=1}^{k-1}\varepsilon_{\tau+i} + \varepsilon_{\tau+k}. \tag{1.5}$$

(By convention, $\sum_{i=1}^{0}\varepsilon_i \equiv 0$). Given knowledge available in period $\tau - 1$, the one-step-ahead forecast for the demand in period τ is $E_{\tau-1}\left[d_\tau\right] = d_{\tau-1} + \mu - (1-\lambda)\varepsilon_{\tau-1}$, so that $\varepsilon_\tau = d_\tau - E_{\tau-1}\left[d_\tau\right]$.

Forecasting may also be done by exponential smoothing, adjusted for the deterministic drift. Let

$$\tilde{d}_t = \lambda d_t + (1-\lambda)\left(\tilde{d}_{t-1} + \mu\right),\ \tilde{d}_0 = 0 \tag{1.6}$$

be the smoothed demand process. Then, working recursively, we find $\tilde{d}_t = \mu t + \lambda \sum_{i=1}^{t} \varepsilon_i, t \geq 1$. Since $\tilde{d}_{t-1} + \mu = E_{t-1}[d_t]$, drift-adjusted exponential smoothing using smoothing parameter λ provides optimal forecasts, a fact first noticed by Muth [13]. From period τ, the future demand in any period $\tau + k$ depends on the past and present only through $d_\tau - (1-\lambda)\varepsilon_\tau = \tilde{d}_\tau$. The smoothed demand process is a discrete random walk with drift μ and variance $(\lambda\sigma)^2$. Let

$$\Delta_k = d_{\tau+k} - \tilde{d}_\tau = \mu k + \lambda \sum_{i=1}^{k-1} \varepsilon_i + \varepsilon_k \qquad (1.7)$$

be the growth in demand over *k* future periods relative to the smoothed demand at a fixed time. The demand growth follows a normal distribution with mean μk and variance $\sigma^2\left(\lambda^2(k-1)+1\right)$.

For large *k*, both the mean and variance of demand growth are (approximately) linear in *k*. We will refer to the coefficient of variation (C.V.) of demand growth as $\lambda\sigma/\mu$. Note that the correlation and randomness parameters combine multiplicatively to yield the overall unpredictability in demand growth. However, over the short term, correlated demand growth can be easier to predict than demand growth that is more random with the same long term C.V.

Let $C_0 > 0$ be the initial capacity. Assume that the cost, $f(x)$, of an expansion of size *x* is an increasing concave function. Because of such economies of scale as well as practical construction and/or installation considerations, capacity will be added in discrete increments rather than continuously over time. An expansion policy is defined by a sequence $\{(t_n, X_n), n \geq 1\}$, where $t_n > t_{n-1}$ is the *n*th expansion epoch and X_n is the size of the *n*th

expansion. Let K_n be the capacity level after *n* expansions have been completed: $K_0 = C_0$ and $K_n = C_0 + \sum_{i=1}^{n} X_i, n \geq 1$. We assume that there is a fixed lead time, *L* periods, required for installing capacity. Let C_t be the capacity available in period *t*. Then $C_t = K_0$ for $t = 0,1,\ldots t_1 + L$ and, for $n \geq 1, C_t = K_n$ for $t = t_n + L + 1, \ldots t_{n+1} + L$. For simplicity, assume the cost $f(X_n)$ is incurred all at once at time t_n. Costs are discounted continuously at rate *r* per period.

In facilities providing services such as communications, electric power, or the steam used for process heat, any excess available cannot be stored nor can unsatisfied demand be backlogged. Frequently, a capacity shortage or disruption in service causes considerable inconvenience, lost productivity and waste. For example, a shortage of steam for heating chemical processes may result in reduced quality or even the disposal of a large quantity of improperly processed material. Rather than trying to assign a monetary penalty to capacity shortages, managers may prefer to fix a service level requirement and then minimize the cost of meeting it. The capacity expansion problem is to choose the expansion epochs and sizes $\{(t_n, X_n), n \geq 1\}$ in order to minimize the infinite horizon expected discounted cost of expansions subject to a maximum allowable expected shortage during any lead time.

3. Form of the Expansion Policy

The service level is an expression of rarity and/or negligibility of any shortages that occur. If the service requirement is high, there should usually be ample capacity to satisfy demand. Given lead times for adding capacity, the timing problem is to determine how far in advance of its need to begin building additional capacity. The only real danger of shortage occurs during an

expansion lead time. In this section we develop a timing criterion. It is possible that lead times could overlap, i.e., $t_n + L > t_{n+1}$ for some *n* if demand grows so quickly during the *n*th lead time that the timing criterion is already satisfied before increment X_n comes online. To avoid counting the same shortages more than once, the timing criterion selects a value of t_n to control the possibility of shortages in periods $\max\{t_{n-1}+L, t_n\}+1, \ldots, t_n + L$. Figure 1 illustrates a portion of an expansion policy for a demand realization in which two lead times overlap.

*** Figure 1 Here ***

Let $g(\mathbf{z})$ be a nondecreasing function of $\mathbf{z} = (z_1, \ldots, z_L)$, i.e., $\mathbf{z}^1 \le \mathbf{z}^2 \Rightarrow g(\mathbf{z}^1) \le g(\mathbf{z}^2)$ where $\mathbf{z}^1 \le \mathbf{z}^2$ if $z_k^1 \le z_k^2$ for each $k = 1, \ldots, L$. The (random) shortage attributed to the *n*th lead time is given by $g(\mathbf{z}^{(n)})$, where

$$z_k^{(n)} = \begin{cases} 0 \text{ if } k \le t_{n-1} + L - t_n \\ \left(d_{t_n+k} - K_{n-1}\right)^+ \text{ if } k > t_{n-1} + L - t_n. \end{cases}$$

For instance, simple shortage measures include the total shortage, $g(\mathbf{z}) = \sum_{k=1}^{L} z_k$; the maximum shortage, $g(\mathbf{z}) = \max\{z_1, \ldots, z_L\}$; or the fraction of periods during the lead time in which a shortage occurs, $g(\mathbf{z}) = \frac{1}{L}\sum_{k=1}^{L} I[z > 0]$, where $I[\cdot]$ is an indicator function that equals 1 if the inequality in brackets is true and 0 otherwise.

From Equation (1.5),

$$d_{t_n+k} - K_{n-1} = \lambda \sum_{i=1}^{k-1} \varepsilon_{t_n+i} + \varepsilon_{t_n+k} + \mu k - \left(K_{n-1} - \tilde{d}_{t_n} \right), \tag{1.8}$$

that is, the shortage in the *n*th lead time depends on decisions and events up to time t_n only through $e_n \equiv K_{n-1} - \tilde{d}_{t_n}$, the capacity position in excess of the smoothed demand at time t_n. Specifically, for $k > t_{n-1} + L - t_n$, $z_k^{(n)}$ has the same distribution as $(\Delta_k - e_n)^+$. A larger value of e_n reduces both the likelihood and expected magnitudes of shortages during the *n*th lead time. If decisions were made in continuous time, a natural timing criterion would specify an exact value for this excess capacity that would trigger an expansion. Our timing policy adapts this idea for periodic observations of demand.

Lemma 1: Suppose $h(s) = E\left[g\left((\Delta_2 - s)^+, \ldots, (\Delta_{L+1} - s)^+ \right) \right]$. Then $h(s)$ is a nonnegative, nondecreasing continuous function with $\lim_{s \to \infty} h(s) = 0$.

Proof: Suppose $s_1 < s_2$. Then for each *k*, $(\Delta_k - s_2)^+ \leq_{st} (\Delta_k - s_1)^+$. Therefore, since *g* is nondecreasing, $h(s_1) \leq h(s_2)$ follows from Definition 1.10.1 of [17]. The continuity of $h(s)$ follows from the fact that $\Delta_2, \ldots, \Delta_{L+1}$ are jointly continuous random variables. As $s \to 0, (\Delta_k - s)^+ \to 0$ with probability 1.

Theorem 1 (Timing Policy): Let $s^*(p) = \min\{s : h(s) \leq p\}$, where $h(s)$ is as defined in Lemma 1. For a fixed *p* > 0, let the *n*th expansion epoch be $t_n = \min\left\{ t : \tilde{d}_t \geq K_{n-1} - s^*(p) \right\}$. Then $E\left[g\left(z_1^{(n)}, \ldots, z_L^{(n)} \right) \right] \leq p$.

Proof: Trivially, $z_k^{(n)} = 0 \leq_{st} \left(\Delta_{k+1} - s^*(p) \right)$ if $k \leq t_{n-1} + L - t_n$. For $k > t_{n-1} + L - t_n$,

$$d_{t_n+k} - K_{n-1} = \tilde{d}_{t_n-1} + \mu(k+1) + \lambda\sum_{i=0}^{k-1}\varepsilon_{t_n+i} + \varepsilon_{t_n+k} - K_{n-1}.$$

From the definition of t_n, $\tilde{d}_{t_n-1} < K_{n-1} - s^*(p)$. Therefore,

$d_{t_n+k} - K_{n-1} < \mu(k+1) + \lambda\sum_{i=0}^{k-1}\varepsilon_{t_n+i} + \varepsilon_{t_n+k} - s^*(p)$, where the quantity on the right hand side is distributed as $\Delta_{k+1} - s^*(p)$. Therefore $z_k^{(n)} \leq_{st} \left(\Delta_{k+1} - s^*(p)\right)^+$ for all k, and since g is nondecreasing, it follows that $E\left[g\left(z_1^{(n)},\ldots,z_L^{(n)}\right)\right] \leq h\left(s^*(p)\right) \leq p$.

Note that t_n could also be defined as $t_n = \min\left\{t : \tilde{D}_t \geq K_{n-1} - s^*(p)\right\}$, where $\tilde{D}_t \equiv \max_{u=1,\ldots,t}\left\{\tilde{d}_u\right\}$ is the maximum demand observed up to time *t*. Also, while we cannot guarantee that lead times will not overlap, we can ensure that $t_n < t_{n+1}$ for all *n* by choosing X_n large enough. Clearly, $t_{n+1} \geq t_n$. If $s = s^*(p)$, the event that $t_{n+1} = t_n$ occurs only if $\tilde{d}_t < K_{n-1} - s$ for all $t < t_n$ and $\tilde{d}_{t_n} > K_n - s$. However, since the error term $\varepsilon_{t'}$ is known at time t_n, we can require that $X_n > \tilde{d}_{t_n} - \tilde{d}_{t_n-1} = \mu + \lambda\varepsilon_{t_n}$, so that $\tilde{d}_{t_n} < \tilde{d}_{t_{n-1}} + X_n < K_n - s$. This requirement is not restrictive when there are expansion economies of scale or technical restrictions on the minimum size of an expansion.

From Theorem 1, a facility manager can specify a maximum tolerable amount of expected shortage during an expansion lead time (specifically, the proportion of the lead time that does not overlap the previous lead time), and find a minimum value of excess capacity that should trigger an expansion. Timing is the primary focus of this paper because it determines the potential lead time shortages and also influences the expansion cost. Analysis of this cost can be simplified by

approximating the smoothed demand, a random walk with drift $\mu > 0$ and variance $(\lambda\sigma)^2$, as a continuous time Brownian motion process with the same drift and variance. Using the continuous time approximation, each expansion epoch has a known distribution in terms of previous expansion sizes. The *n*th expansion epoch, t_n, is approximately equal to $T\left(K_{n-1} - s\right)$, where $s = s^*\left(p\right)$ for some specified *p*, and $T\left(a\right)$ is the first time the Brownian motion process reaches the value *a*. By a well-known result (see, for example, Karlin and Taylor [10]),

$$E\left[e^{-rT(a)}\right] = e^{-\rho a}, \text{ where } \rho = \frac{\sqrt{\mu^2 + 2r\left(\lambda\sigma\right)^2} - \mu}{\left(\lambda\sigma\right)^2}.$$

Therefore, under the timing policy, for a given sequence of expansion sizes, the infinite horizon expected discounted expansion cost is given by

$$F\left(s, \left\{X_n, n \geq 1\right\}\right) = \sum_{n=1}^{\infty} E\left[e^{-rt_n}\right] f\left(X_n\right) = \sum_{n=1}^{\infty} e^{-\rho\left(K_{n-1} - s\right)} f\left(X_n\right).$$

First, we show that an optimal size policy exists.

Lemma 2: If $X_n = X > 0$ for all *n* = 1, 2, …, then $F\left(s, \left\{X_n, n \geq 1\right\}\right) < \infty$.

Proof:

$$F\left(s, \left\{X, n \geq 1\right\}\right) = e^{-\rho\left(C_0 - s\right)} \sum_{n=1}^{\infty} e^{-\rho(n-1)X} f\left(X\right) = e^{-\rho\left(C_0 - s\right)} f\left(X\right) \sum_{n=0}^{\infty} \left(e^{-\rho X}\right)^n = \frac{e^{-\rho\left(C_0 - s\right)} f\left(X\right)}{1 - e^{-\rho X}} < \infty.$$

Next, a dynamic programming argument shows that there is an optimal policy with equal expansion sizes.

Theorem 2 (Size Policy): Under the timing policy, there is an optimal sequence of expansion sizes $\left\{X_n, n \geq 1\right\}$ with $X_n = X$ for all *n*.

Proof: Suppose $\{X_n, n \geq 1\}$ minimizes

$e^{\rho(C_0 - s)} F\left(s, \{X_n, n \geq 1\}\right) = f\left(X_1\right) + e^{-\rho X_1}\left(\sum_{n=2}^{\infty} \exp\left(-\rho \sum_{i=2}^{n-1} X_i\right) f\left(X_n\right)\right)$. Then, by the principle of optimality, $\{X_n, n \geq 2\}$ minimizes $\sum_{n=2}^{\infty} \exp\left(-\rho \sum_{i=2}^{n-1} X_i\right) f\left(X_n\right)$, i.e., the sequence $X_2 = X_1, X_3 = X_2, \ldots$ is optimal from time t_2 onward. Repeated applications of this argument yield the result.

The expansion policy developed here is consistent with previous results that were obtained without lead times for the cases when $\lambda = 0$ or $\lambda = 1$. Manne [11] and Giglio [8] assumed demand could be added instantaneously but allowed the possibility of planned shortages. These previous analyses can be viewed as shifting the expansion cycles so that a (zero-lead-time) expansion epoch corresponds to the end of a lead time in our analysis. For Manne's demand model, a continuous time version of the smoothed IMA process with $\lambda = 1$, it is optimal to expand capacity when demand exceeds capacity by some fixed amount. In Giglio's model, when demand follows a linear trend process ($\lambda = 0$), the optimal policy expands capacity when demand differs from *expected* capacity by some fixed amount. Our timing policy can be restated in terms of actual, not smoothed demand, as $t_n = \min\left\{t : d_t \geq K_{n-1} - s^*\left(p\right) + \left(1 - \lambda\right)\varepsilon_t\right\}$. The excess capacity threshold that triggers an expansion is adjusted by the deviation in the current demand from what was expected. If $\lambda = 1$ the actual demand is used as in [11]. If $\lambda = 0$, timing decisions are made according to the expected demand as in [8]. Since, under the timing policy, the future appears identical at each expansion epoch, we obtain constant optimal expansion sizes in a manner consistent with both these papers as well as [5]. The value of X^* that minimizes

$F\left(s,\{X,n\geq 1\}\right)$ is independent of $s = s^*(p)$. However, the value of the optimal cost is influenced by both $s^*(p)$ and X^*. In the next section, we examine the effects of correlation and randomness on the optimal policy parameters and the total cost.

4. Effects of Correlation and Randomness on Threshold

For $k \geq 1$, let $\Delta_k(\lambda,\sigma) = \mu k + \lambda\sum_{i=1}^{k}\varepsilon_{\tau+i} + \varepsilon_{\tau+k}$, where the $\varepsilon_1,\ldots,\varepsilon_k$ are independent, identically distributed normal random variables with mean 0 and variance σ^2. We can show that, for a given level of excess capacity, the shortage in any period in the lead time is stochastically increasing in the variance of $\Delta_k(\lambda,\sigma)$. Therefore, for certain measures of total lead time shortage, the expected shortage is also monotonic in the variance of $\Delta_k(\lambda,\sigma)$.

Lemma 3. Let (λ_1,σ_1) and (λ_2,σ_2) satisfy $\sigma_1^2\left(\lambda_1^2(k-1)+1\right) < \sigma_2^2\left(\lambda_2^2(k-1)+1\right)$. Then for $k \geq 1$ and any $s \geq 0$, $\left(\Delta_k(\lambda_1,\sigma_1)-s\right)^+ \leq_{st} \left(\Delta_k(\lambda_2,\sigma_2)-s\right)^+$.

Proof: Fix $k \geq 1$. For i = 1, 2, let $F_i(x) = \Pr\left[\left(\Delta_k(\lambda_i,\sigma_i)-s\right)^+ \leq x\right]$. For $x \leq 0$, $F_1(x) = F_2(x) = 0$. For $x > 0$,

$$F_1(x) = \Pr\left[\Delta_k(\lambda_1,\sigma_1) \leq s+x\right]$$
$$= \Phi\left[\frac{s+x-\mu k}{\sigma_1\sqrt{\lambda_1^2(k-1)+1}}\right] > \Phi\left[\frac{s+x-\mu k}{\sigma_2\sqrt{\lambda_2^2(k-1)+1}}\right] = F_2(x),$$

where $\Phi[\cdot]$ is the standard normal cumulative distribution function. Therefore, $F_1(x) \geq F_2(x), \forall x$.

The condition of Lemma 3 holds in particular for (λ_1, σ) and (λ_2, σ) with $\lambda_1 < \lambda_2$, for (λ, σ_1) and (λ, σ_2) with $\sigma_1 < \sigma_2$, and for $(\lambda_1, CV/\lambda_1)$ and $(\lambda_2, CV/\lambda_2)$ with $\lambda_1 > \lambda_2$. Its important implication is that the excess capacity threshold used in the timing policy increases with the variance of demand growth.

Theorem 3. Let (λ_1, σ_1) and (λ_2, σ_2) satisfy the condition of Lemma 3. For *i* = 1, 2, let $h_i(s) = E\left[g\left(\left(\Delta_2(\lambda_i, \sigma_i) - s\right)^+, \ldots, \left(\Delta_{L+1}(\lambda_i, \sigma_i) - s\right)^+ \right) \right]$, where $g(\cdot)$ is nondecreasing, and let $s_i^*(p) = \min\{s : h_i(s) \le p\}$. Then $s_1^*(p) \le s_2^*(p)$.

Proof: As in the proof of Lemma 1, $h_1(s) \le h_2(s)$ for all *s*. The result is immediate.

Although certain measures of expected lead time shortage have attractive properties, actually estimating the shortage in order to compute the capacity threshold level is not easy. Also, insisting on measuring shortages using simple nondecreasing functions of the shortages in each period may be too restrictive. We used simulation to estimate lead time shortages that would result from applying various values of the threshold level, *s*, in the timing policy. Taking advantage of the flexibility of simulation, we evaluated shortages according to the proportion of demand growth during the lead time that is not satisfied. Assuming for simplicity that excess capacity exactly equals *s* when an expansion begins, this shortage measure is given by: $G(s) = \sum_{k=1}^{L} (\Delta_k - s)^+ \Big/ \sum_{k=1}^{L} (\Delta_k)^+$. For a given *p*, we estimated $s^*(p) = \min\{s : E[G(s)] \le p\}$. Because the variation in demand growth affects the denominator as well as the numerator, it is not so clear that it causes *G*(*s*) to stochastically increase. However, one would expect the impact of increasing demand variance on the numerator to exceed that on the denominator as the

numerator is smaller. In the numerical simulation results, we verified that the average value of $G(s)$ over replications had the same properties as are proved for $h(s)$ in Lemmas 1 and 3. Therefore, the definition of $s^*(p)$ in this context makes sense, and the qualitative effects of λ and σ agree with Theorem 3. The simulation results quantify and contrast these effects on the capacity threshold.

Two simulation experiments were performed to determine how demand and lead time characteristics affect the excess capacity threshold for a given service level. The first looked at separate effects of σ, λ, and L. The second compared more closely the relative sizes and types of the impacts of randomness and correlation.

4.1. Separate Effects

The first simulation experiment was run to determine the effects of σ, λ, and L on

$$s^*(p) \equiv \min\left\{s : E\left[G(s)\right] \le p\right\} \tag{1.9}$$

for a given value of p. Demand realization i yielded a sample $G_i(s)$ for discrete values of s up to a quantity S equal to the 99.9% upper prediction limit for demand growth during a lead time. An estimate of $s^*(p)$ was obtained using the average of $\{G_i(s)\}$ over the realizations. This process was replicated a number of times to obtain the mean and standard deviation of $s^*(p)$. The algorithm proceeded as follows:

Algorithm 1

1. For each combination of σ, λ, and L, do Steps 2 - 10

2. Compute $S = \mu L + z_{0.001}\sigma\sqrt{1+(L-1)\lambda^2}$, $\delta = S/N$, and let $s[n] = n\delta, n = 1,\ldots N$.

3. For j = 1, …, J do Steps 4 - 9

   4. For i = 1, …, I do Steps 5 - 8

      5. Initialize $d_0 - \theta\varepsilon_0 = 0$, $u_i(s[n]) = 0$, and $v_i = 0$, where $u_i(s[n])$ represents the unmet demand for $s[n]$ in the i^{th} demand realization and v_i is the cumulative (positive) demand growth in realization i.

      6. For t = 1,…, L do Step 7

         7. Randomly generate ε_t and obtain d_t from Equation (1.1). Set $v_i = v_i + (d_t)^+$. For n = 1, … , N, set $u_i(s[n]) = u_i(s[n]) + (d_t - s[n])^+$.

      (end loop for t)

      8. Compute $G_i(s[n]) = u_i(s[n])/v_i$, for n = 1, … , N.

      (end loop for i)

   9. Compute $\bar{G}(s[n]) = \frac{1}{M}\sum_{i=1}^{M} G_i(s[n])$, for n = 1, … , N. Find $s_j^*(p) \equiv \min\{s[n] : \bar{G}(s[n]) < p\}$.

   (end loop for j)

10. Compute the mean and standard deviation of $\{s_j^*(p), j = 1,\ldots, J\}$.

(end loop for σ, λ, and L).

We scaled the demand growth by setting μ = 1. The parameter values tested were chosen to reflect a situation of significant lead times as well as demand uncertainty and correlation:

$L \in \{12, 24, 48\}$, $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$, $\sigma \in \{1, 2, 3, 4\}$. We tested N = 100 values of s over M = 50 demand realizations and averaged the s^* values over J = 50 replications. Finally, we set p = 0.01 to find a value of s that would meet 99% of demand growth during the expansion lead time.

The effects on the mean $s^*(0.01)$ of σ, λ and L are shown in Figures 2, 3 and 4, respectively, which represent three different views of the same numerical results. The widths of the 95% confidence intervals were less than 4% of the mean in all cases. Figures 2 and 3 confirm that correlation acts similarly to randomness to prompt expansions earlier, i.e., when more excess capacity remains. The interaction between these two factors is seen in the fact that the effect of increasing σ is larger when λ is larger and vice versa.

*** Figures 2, 3, and 4 Here ***

Figure 4 shows that longer lead times also result in earlier expansions. For comparison, the chart also includes the line for L, the expected demand growth during the lead time with μ = 1. When σ and/or λ are large, the excess capacity threshold is significantly larger than L. When both these parameters are small, the values of s^* determined for larger values of L are slightly smaller than the expected lead time demand growth.

4.2. Comparative Effects of Nonstationarity and Randomness

The first simulation results indicate that correlation can affect the choice of s^* in a similar manner as randomness, particularly for longer lead times. Each of these contributes to the unpredictability of demand: correlation, in that the mean demand growth fluctuates over time;

and randomness, in that demand varies about its mean. Equations (1.3) and (1.4) show that if C.V. = $\lambda'\sigma' = \lambda''\sigma''$ with $\lambda' < \lambda''$, then the variance of demand in any period is larger for (λ', σ'), while the correlation between demands in nearly consecutive periods is larger for (λ'', σ'').

To explore the relative effects of correlation and randomness numerically, we designed a simulation experiment with two different (λ, σ) pairs for the same C.V. as follows.

Algorithm 2

1. For each C.V. do Steps 2 - 13
    2. Compute $S = \mu L + z_{0.001}\sigma\sqrt{1+(L-1)\lambda^2}$, $\delta = S/N$, and let $s[n] = n\delta, n = 1, \ldots N$.
    3. For j = 1, …, J do Steps 4 - 12
        4. For i = 1, …, M, let $R[i]$ be a randomly generated seed.
        5. For each of two pairs (λ', σ') and (λ'', σ'') such that $\lambda'\sigma' = \lambda''\sigma'' =$ C.V. with $\lambda' < \lambda''$, do
            6. For i = 1, …, M do Steps 7 - 10
                7. Initialize $d_0 - \theta\varepsilon_0 = 0$, $u_i(s[n]) = 0$, and $v_i = 0$. Seed the random number generator with $R[i]$.
                8. For t = 1,…, L do Step 9
                    9. Randomly generate ε_t and obtain d_t from Equation (1.1). Update v_i and $u_i(s[n])$ as in *Algorithm 1*.

                (end loop for t)

10. Compute $G_i(s[n]) = u_i(s[n])/v_i$, for n = 1, … , N.

(end loop for i)

11. Compute $\bar{G}(s[n]) = \frac{1}{M}\sum_{i=1}^{M} G_i(s[n])$, for n = 1, … , N. Find $s_j^*(p)$ and record as s_j' or s_j'' as appropriate for (λ', σ') or (λ'', σ''), respectively.

(end loop for (λ', σ') and (λ'', σ''))

12. Compute $s_j^d = s_j' - s_j''$.

(end loop for j)

13. Compute the mean and standard deviation of $\{s_j^d, j = 1, \ldots, J\}$.

(end loop for C.V.).

We tested C.V.'s within the range 0.5 to 1.5 and pairs (λ, σ) with $\sigma' = 2\sigma'', \sigma'' = 1$ or 2. Figure 5 plots the mean difference of $s^*(0.01)$ for higher σ less $s^*(0.01)$ for lower σ. All the means are significantly greater than zero, with 95% confidence interval widths ranging from 0.03 to 0.10. The fact that all the differences are positive indicates that randomness is worse than correlation, in that earlier capacity additions are required to meet the same service level criterion for demand parameters (λ', σ') than for (λ'', σ''). The difference is larger when the long term C.V. is smaller and, for the same C.V., when both σ values are larger. The effects of correlation and randomness on timing differ more when lead times are shorter.

*** Figure 5 Here ***

These results highlight the importance for timing expansions of specifying the demand process correctly and estimating its parameters accurately. In the next section we show that if correlation is ignored or specified incorrectly, the amount of randomness in the demand will be overestimated. As a result, managers may act over-cautiously to expand capacity well before demand reaches the current capacity.

5. Impact of Errors in Specifying and Estimating the Demand Process

Given a historical pattern for demand that generally appears to be linearly increasing, a decision maker may fail to recognize correlation or mis-estimate its extent. If the demand process is assumed to be stationary, then the mean and variance of demand growth are estimated according to a simple linear trend model:

$$d_t = \mu t + \varepsilon_t . \tag{1.10}$$

Estimates $\hat{\mu}$ and $\hat{\sigma}$ of the parameters μ and σ can be estimated by linear regression with *t*. Alternatively, the analyst may acknowledge correlation that leads to nonstationarity in demand by applying exponential smoothing but choose an incorrect smoothing parameter, $0 < \hat{\lambda} \neq \lambda \leq 1$. If $\hat{\lambda} = 1$ the process is viewed as a pure random walk, and the "smoothed" demand equals the current demand. In order to focus on correlation, we assume that the linear trend parameter (μ) is estimated correctly, i.e., $\hat{\mu} = \mu$, and examine the effect of ignoring or choosing an incorrect value for λ.

To ignore or mis-estimate λ is to poorly fit a demand process to the historical demand and, therefore, poorly forecast future demand. First consider correctly specified demand model with a

bad guess for λ. Correcting for a trend in the demand and then performing exponential smoothing is equivalent to assuming that demand follows the IMA process:

$$d_t = d_{t-1} + \mu - \left(1-\hat{\lambda}\right)\varepsilon_{t-1} + \varepsilon_t . \tag{1.11}$$

The random error term ε_t is estimated as the deviation of d_t from the one-step-ahead forecast of demand in the previous period, given by $E_{t-1}[d_t] = d_{t-1} + \mu - \left(1-\hat{\lambda}\right)\varepsilon_{t-1}$. However, in reality, the demand follows the process given by Equation (1.1). Based on data up to time *T*, the variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{T-2}\sum_{t=1}^{T}\left(d_t - E_{t-1}[d_t]\right)^2 = \frac{1}{T-2}\sum_{t=1}^{T}\left(\varepsilon_t + \left(\lambda-\hat{\lambda}\right)\varepsilon_{t-1}\right)^2 ,$$

where the sum of squared deviations is divided by *T* - 2 since two parameters, $\hat{\mu}$ and $\hat{\lambda}$, have been estimated. Therefore,

$$E\left[\hat{\sigma}^2\right] = \frac{1}{T-2}\left[T\sigma^2 + (T-1)\left(\lambda-\hat{\lambda}\right)^2\sigma^2\right] \underset{T\to\infty}{\rightarrow} \sigma^2\left[1+\left(\lambda-\hat{\lambda}\right)^2\right] > \sigma^2 . \tag{1.12}$$

From Lemma 3, we know that the shortage in the *k*th period of a lead time stochastically increases with the quantity $\sigma^2\left(\lambda^2(k-1)+1\right)$. Clearly, if $\hat{\lambda} > \lambda$, then inequality (1.12) implies that $E\left[\hat{\sigma}^2\left(\hat{\lambda}^2(k-1)+1\right)\right] > \sigma^2\left(\lambda^2(k-1)+1\right)$. For a specified value of *p*, the effect of overestimating the correlation parameter is to choose a larger value of $s^*(p)$, so that expansions would occur earlier. On the other hand, if $\hat{\lambda} < \lambda$, one can show that $E\left[\hat{\sigma}^2\left(\hat{\lambda}^2(k-1)+1\right)\right] < \sigma^2\left(\lambda^2(k-1)+1\right)$. The expected shortage during the lead time will be underestimated and expansions occur too late.

Errors in estimating the demand parameters also affect the size of expansions but this effect is less significant than the effect on timing. For a chosen *s*, the optimal expansion size, X^*, is the argument that minimizes $K(X) \equiv f(X) \big/ \left(1 - e^{-\rho X}\right)$. Since the adjusted interest rate, ρ, depends on λ and σ, the value of X^* found according to $\hat{\lambda}$ and $\hat{\sigma}$ will differ from that corresponding to the true λ and σ. However, Manne [11] showed that $K(X)$ is relatively insensitive to *X*. Under our expansion policy, the infinite horizon cost $F(s, X) = e^{-\rho(C_0 - s)} K(X)$ depends on *s* as well as *X*. Let $\left(\hat{s}^*, \hat{X}^*\right)$ be the optimal policy parameters found according to $\hat{\lambda}$ and $\hat{\sigma}$ while $\left(s^*, X^*\right)$ are the optimal parameters for the true demand parameters. For the parameter values used in the simulation, we have verified numerically that, when $f(X) = X^{0.7}$,

$\left|F\left(\hat{s}^*, \hat{X}^*\right) - F\left(s^*, X^*\right)\right| \cong \left|F\left(\hat{s}^*, X^*\right) - F\left(s^*, X^*\right)\right| >> \left|F\left(s^*, \hat{X}^*\right) - F\left(s^*, X^*\right)\right|$. Therefore, the timing of expansions controls most of the expansion cost (as well as all of the penalty due to shortages).

Finally, consider the effect of ignoring correlation completely ($\hat{\lambda} = 0$). In this case, the most natural forecasting method for the process in Equation (1.10) is to estimate d_0 and μ by simple linear regression with *t* as the independent variable. Based on *T* observations of past demand, the estimate for σ^2 is given by

$$\hat{\sigma}^2(T) = \frac{1}{T-2} \sum_{t=1}^{T} \left(d_t - \hat{d}_0 - \hat{\mu} t\right).$$

Then, assuming $\hat{d}_0 = d_0$ and $\hat{\mu} = \mu$, the expected value of this estimate is

$$E\left[\hat{\sigma}^2(T)\right] = E\left[\frac{1}{T-2}\sum_{t=1}^{T}\left(\lambda\sum_{i=1}^{t-1}\varepsilon_i + \varepsilon_t\right)\right] = \sigma^2\,\frac{\lambda^2T^2 + \left(2-\lambda^2\right)T}{2T-4}.$$

Note that $\hat{\sigma}^2(T) \to \infty$ as $T \to \infty$.

Giglio [8] approximated optimal policies assuming zero lead times, but allowing the possibility of planned shortages. (He referred to this process as "nonstationary" due to the presence of a linear trend.) He argued that, relative to deterministic linear demand, the optimal (constant) expansion size is unaffected by the presence of uncertainty, but when shortage costs are high, increasing uncertainty in demand provokes expansions earlier, i.e., when higher excess capacity remains. Therefore, the mis-specification of demand as a linear trend process results in an overestimate of the demand variance, a higher reserve margin of excess capacity, and an increase in the discounted expansion cost due to earlier expansions. These effects are exacerbated as more data are used in estimating demand parameters.

6. Conclusions

The choice of an expansion policy is complicated by lead times for adding capacity combined with autocorrelation that can lead to nonstationary demand growth. However, if the demand correlation is treated appropriately, the optimal policy has a form that is consistent with classical results for capacity expansion. This paper shows that correlation in demand has a significant effect on the timing of expansions but that its effect is less than that of randomness, particularly when lead times are short. When shortages are expensive, the sizes of expansions are generally less important than their timing.

The use of common forecasting techniques such as exponential smoothing and ARIMA involves an implicit assumption that demand is correlated. However, demand correlation has not been considered explicitly in most previous expansion models. Instead, authors have assumed demand follows either a linear trend process or a pure random walk. Failing to account for demand correlation properly can have serious consequences for the expansion policy and its long term cost.

Acknowledgment: This work was supported by the National Science Foundation under grant DMI-9996373. The simulations were carried out by Suzanne Childress under a Research Experiences for Undergraduates supplement.

References

[1] A. Angelus, E. Porteus, S.C. Wood, Optimal sizing and timing of capacity expansions with implications for modular semiconductor wafer fabs. In:, Graduate School of Business, Stanford University, Stanford, CA, 1997.

[2] J.C. Bean, J. Higle, R.L. Smith, Capacity expansion under stochastic demands, Operations Research 40 (1992) S210-S216.

[3] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[4] M. Çakanyildirim, R.O. Roundy, S.C. Wood, Optimal capacity expansion strategies under demand uncertainty. In:, Cornell University, Ithaca, NY, 2001.

[5] B.A. Chaouch, J.A. Buzacott, The effects of lead time on plant timing and size, Production and Operations Management 3 (1994) 38-54.

[6] M.H.A. Davis, M.A.H. Dempster, S.P. Sethi, D. Vermes, Optimal capacity expansion under uncertainty, Advances in Applied Probability 19 (1987) 156-176.

[7] J. Freidenfelds, Capacity Expansion: Analysis of Simple Models with Applications, North-Holland, New York, 1981.

[8] R.J. Giglio, Stochastic capacity models, Management Science 17 (1970) 174-184.

[9] S.C. Graves, A single-item inventory model for a nonstationary demand process, Manufacturing & Service Operations Management 1 (1999) 50-61.

[10] S. Karlin, H.M. Taylor, A First Course in Stochastic Processes, Academic Press, New York, 1975.

[11] A.S. Manne, Capacity expansion and probabilistic growth, Econometrica 29 (1961) 632-649.

[12] C.D. McAllister, S.M. Ryan, Relative risk characteristics of rolling horizon hedging heuristics for capacity expansion, The Engineering Economist 45 (2000) 115-128.

[13] J.F. Muth, Optimal properties of exponentially weighted forecasts, Journal of the American Statistical Association 55 (1960) 299-306.

[14] S. Nickell, Uncertainty and lags in the investment decisions of firms, Review of Economic Studies 44 (1977) 249-263.

[15] S.M. Rocklin, A. Kashper, G.C. Varvaloucas, Capacity expansion/contraction of a facility with demand augmentation dynamics, Operations Research 32 (1984) 133-147.

[16] S.M. Ryan, Forecast frequency in rolling horizon hedging heuristics for capacity expansion, European Journal of Operational Research 109 (1998) 550-558.

[17] D. Stoyan, Comparison Methods for Queues and Other Stochastic Models, John Wiley & Sons, New York, 1983, 217 pp.

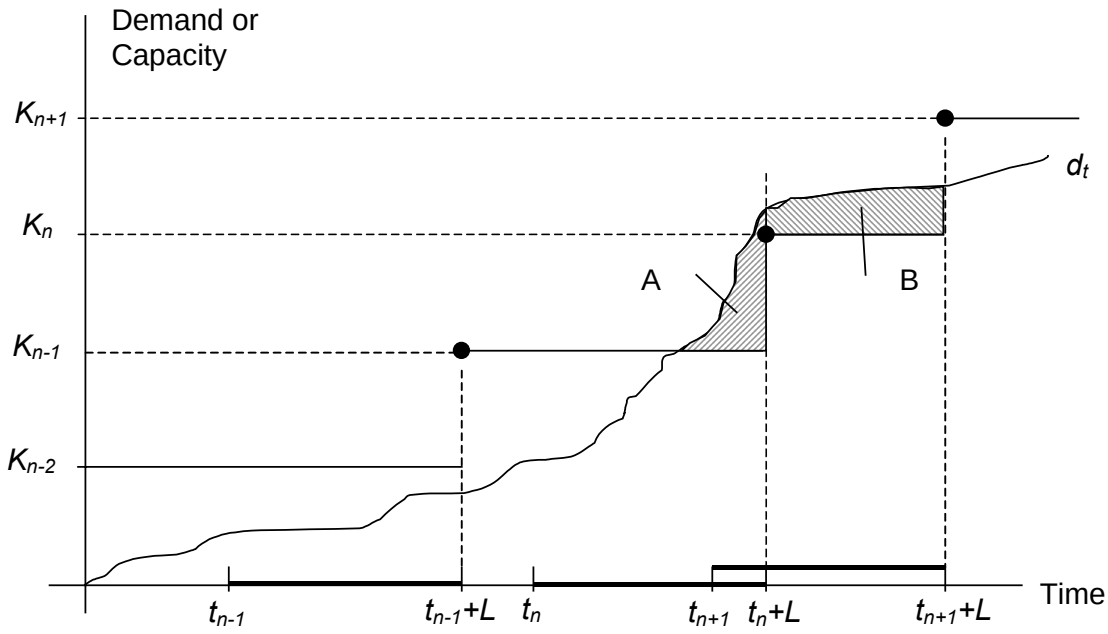

Figure 1. Illustration of capacity expansion policy and the allocation of shortages between overlapping lead times. Shortage A (B) is attributed to the *n*th (resp. (*n*+1)st) lead time.

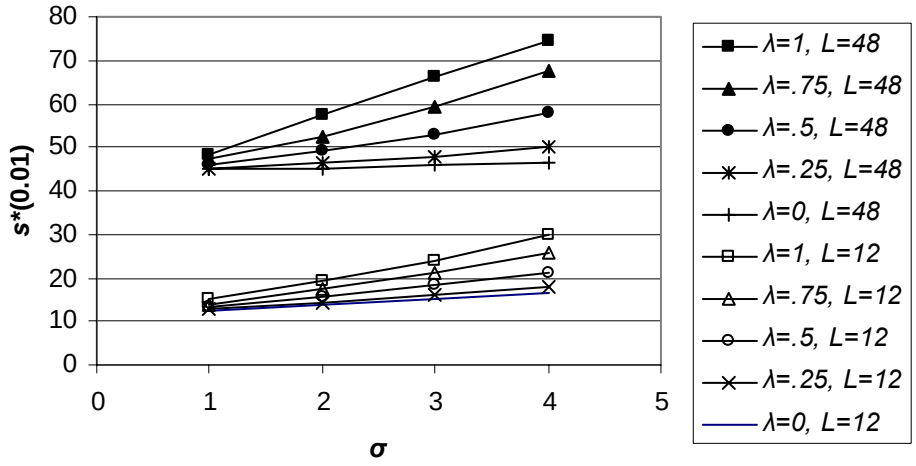

Figure 2. Excess capacity threshold, in capacity units, vs. randomness in demand.

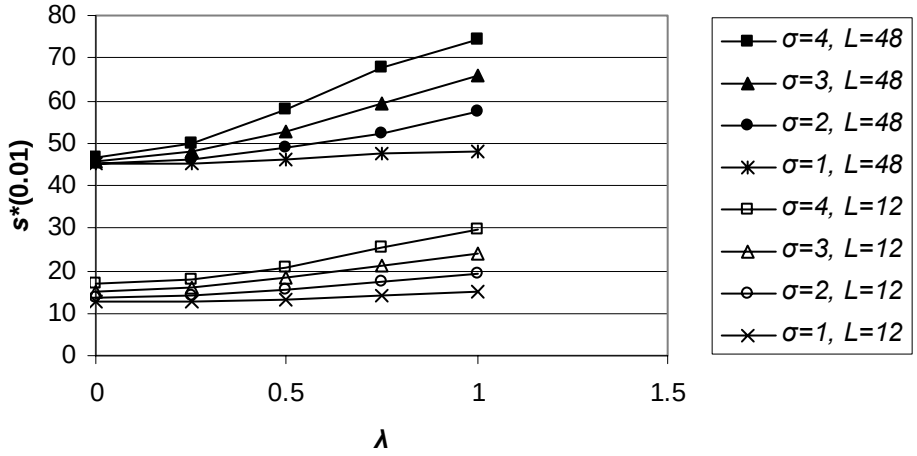

Figure 3. Excess capacity threshold, in capacity units, vs. correlation in demand.

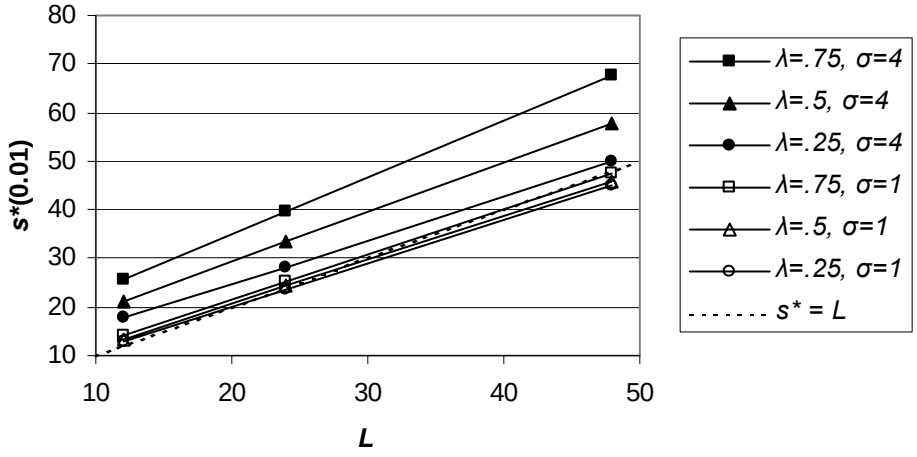

Figure 4. Excess capacity threshold, in capacity units, vs. lead time length.

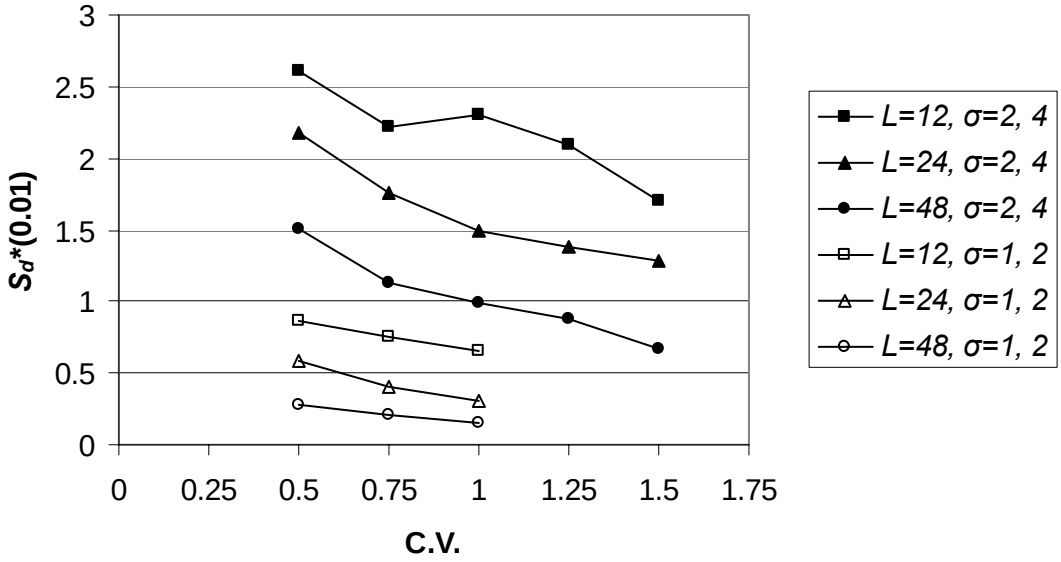

Figure 5. Excess capacity threshold for higher randomness less that for lower randomness for the same C.V.